DOCUMENT RESUME

ED 233 344 CS 207 603

AUTHOR Alexander, John; Swartz, Fred
TITLE The Dynamics of Computer-Assisted Writing Sample Measurement at Ferris State College.
PUB DATE Oct 82
NOTE 17p.; Paper presented at the Annual Meeting of the Michigan Council of Teachers of English (30th, East Lansing, MI, October 29-31, 1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; College Freshmen; *Computer Assisted Testing; Criterion Referenced Tests; Higher Education; *Measurement Techniques; *Student Placement; Test Construction; *Testing Problems; Test Reviews; *Writing (Composition); *Writing Evaluation; Writing Skills
IDENTIFIERS *Computer Assisted Placement; Computer Managed Information

ABSTRACT

Concerned with the failure of multiple choice tests to measure composition--as opposed to grammar and usage--skills, college researchers studied the value of using writing samples to place students in composition classes. Evaluating the writing samples of 300 incoming freshmen, they did not find a high correlation between sample scores and subsequent achievement in courses. Believing that the lack of correlation possibly signaled evaluators' difficulties in maintaining a consistent performance, researchers replaced the more holistic marking of samples with grading based on an analytic, criterion referenced model. To further minimize problems inherent in large-scale writing assessment, they instituted a computer-assisted monitoring system that checked evaluators for agreement, adjusted student scores when divergence occurred in raters' responses, and provided placement recommendations based on the correlation of analytic raw scores and national test scores. Tracking 1,300 tested students as they moved through the freshman program, researchers found that the system minimized the chance for student misplacement, but, as all but 7% of the 565 students entering the basic composition course would have been placed there on the basis of their ACT or SAT score alone, they could not, as yet, give full endorsement to the system. (MM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233344

THE DYNAMICS OF COMPUTER-ASSISTED WRITING SAMPLE MEASUREMENT AT FERRIS STATE COLLEGE

John Alexander
Fred Swartz

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
John Alexander

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS 207603

John Alexander, Head, Department
of Languages and Literature
Fred Swartz, Coordinator of Testing

THE DYNAMICS OF COMPUTER-ASSISTED WRITING SAMPLE MEASUREMENT AT FERRIS STATE COLLEGE

In recent years the work of Paul Dietrich, Lee Odell, and Hunter Breland of the Educational Testing Service has offered evidence that highly reliable and valid measurements of writing ability can be attained by nearly all departments of English that wish to have a role to play in instructional management decisions. Like most English faculty, members of the Department of Languages and Literature at Ferris State College questioned the credibility of multiple-choice assessments of writing performance which up to 1981 were the only examinations used for course placement. Faculty generally felt that such tests, while adequately measuring problems in grammar and usage, and usually only at the sentence level, did not attempt to test the composing process, including the development of and organization of an idea. Even though we were aware of the research that demonstrates that there is no evidence that writing samples provide more reliable measures of complex writing tasks than objective tests,[1] we felt that the challenges outweighed the potential risks. Given the expenditure of time and financial resources necessary for a comprehensive testing program, we thought goals could be met beyond those of accuracy in course placement.

In order to justify the complexity and cost of setting up a comprehensive testing program on a campus of 11,000 students, we were aware that the procedures used should complement broader curricular goals. By offering a writing sample to approximately 3,000 incoming freshmen during the orientation period each summer, we wished to convey the message to students, parents, and our principal feeder schools, that writing and not just correct usage should become the criteria for writing performance. For the faculty in the Department of Languages and Literature, we wished to open avenues of communication concerning composition evaluation and standards. Through group training sessions, we hoped that faculty might arrive at some common understandings concerning

evaluation criteria, thereby improving instruction in the freshman sequence. By using the Dietrich criterion-referenced model for evaluation, we wished to provide an early diagnosis of writing ability, with such a record being sent to the student's orientation advisor, and to the English Department office. The holistic model, used by the ETS, while offering high reliability, low cost, and speed of placement decision, offered little in the way of diagnostic information; consequently, the analytic model, while being more difficult to administer, offered us the most expedient evaluation tool given our goals for the program.

The task of developing a writing sample assessment was far more complex than anyone imagined at the onset of the project. Our preliminary research suggested a number of problems that might be encountered in establishing such a program. Some difficulties were addressed directly, others were beyond our control, and were considered endemic to writing sample administration. We knew from our investigation that a number of variables in the scoring sessions could seriously affect interrater agreement, thereby throwing into question the validity of the placement decision.

Our review of the research confirmed that a writing sample should not stand alone. Despite the fact that standardized, objective tests only measure limited writing skills, studies by Hunter Breland and others show that by adding a writing sample to such a test, a predictive correlation increase of only .01 to .06 can be realized.[2] The problems associated with writing sample measurement were illustrated by statistics generated from an experimental study involving 300 incoming freshmen who were tested with a writing sample in the summer of 1981. While the data is not conclusive since we were unable to track all 300 students, the results tended to support our early concerns. Following is the data correlating the pre-measures (ACT English and Summer Writing Sample) with various post-measures.

		Pre-Measure	Post-Measure	Correlation	Added Contribution of Weaker Correlate
A.	1	Writing Sample	Grades	.163	.013
	2	ACT English	Grades	.375	
	3	Combination	Grades	.382	
B.	1	Writing Sample	Obj. Test Final	.504	.001
	2	ACT English	Obj. Test Final	.806	
	3	Combination	Obj. Test Final	.807	
C.	1	Writing Sample	Writing Sample	.438	
	2	ACT English	Writing Sample	.314	.005
	3	Combination	Writing Sample	.443	
D.	1	Writing Sample	Composite	.567	.013
	2	ACT English	Composite	.752	
	3	Combination	Composite (WS+OT+GR)	.765	

Students finishing their first course in composition were asked to write a writing sample replicating the pre-measure and take an internally designed objective test measuring proofreading ability, usage, and grammar. As block A illustrates, the correlation of grades with the writing sample (.163), ACT English (.375), and their combination (.382) was not high. Furthermore, as demonstrated in Section A, B, and D, the weaker correlate does not add appreciably to the predictive success of the stronger measure. As might be expected, the correlation of the ACT English with the objective departmental (Section B) test was high. The better correlation earned by the writing sample in Section C in predicting success on the writing sample administered at the end of the quarter suggests that "writing samples predict writing samples" in the same way that "grammar tests predict grammar tests." Neither, as evidenced in Section A does a very good job in correlating with grades which may say that success as measured by grades in the first course is determined by variables outside of the scope of our existing measurement knowledge. Using a composite post-measure, combining the final writing sample with an objective test and the final grade, the correlations with the initial writing sample and the ACT are somewhat higher. Yet, as the data shows, the writing sample adds little to the predictive ability of the ACT English test.

The difficulty of maintaining high predictive accuracy in course placement can be due to rater disagreement or to conflicting goals in course objectives. We assumed that by launching a testing program and by carefully monitoring rater performance, we could indirectly affect course goals in the freshman program. But we were also well aware of the difficulty of maintaining a high consistency in rater performance. Our research alerted us to some of the major problems in insuring high agreement.

1. Average quality essays receive higher ratings when preceded by a block of poor essays than when they are preceded by a block of good essays.[3]
2. The quality and frequency of the training sessions have a direct relation to rater performance. Two well-trained raters with 3 writing samples per student perform as dependably as two lesser-trained raters with 6 samples per person.[4] Without periodic training sessions raters will tend to disagree in proportion to the length of the evaluation sessions.
3. Handwriting quality influences raters, especially when the raters expect low achievement.[5]
4. A carefully controlled environment is mandatory if raters are to maintain high levels of agreement. Distractions such as ringing telephones, random chatter, comments from the coordinator about rater performance, all have the effect of diminishing rater performance. If a rater perceives that he is behind other raters, in an effort to "catch up," his scores will deviate from prior performance and the ratings of others on the team.
5. Some papers seem to encourage rater agreement while others produce rater discrepancy. Low interrater agreement and reliability could be caused by the design of the writing assignment, such as ambiguous instructions given to the examinees. The assignment should be as explicit as possible and students should be taken step by step through the process of prewriting

writing, and editing in order to minimize ambiguity in response. As far as possible the design of the assessment should allow students to respond in the same way as they would if they were to receive a take-home assignment.

The choice of the analytic, criterion referenced model for evaluation presented us with other difficulties that threw into question the predictive accuracy of our placement decisions. The analytic assessment procedure involved six scales, each ranked high, middle, and low, with each paper evaluated by two raters. A maximum of 36 points were possible for each reading. Wide score discrepancies, involving 7 or more points, were to be resolved by a third reader, either by the arbiter-coordinator, or by a third rater. The analytic method, while rhetorically less adequate than the primary trait model used by the NAEP, could be quickly implemented with a minimum of training expense. Faculty teams were responsible for defining what is to be measured, developing the scoring criteria, and providing training sessions for new raters. The method allowed the faculty an opportunity to discuss and to work toward common evaluative procedures that would be somewhat consistent with the actual practice in over 100 sections of composition each term. It was our hope that if raters could consistently adhere to the agreed-upon criteria, then some progress could be achieved in reaching a consensus on desirable objectives for each course in the freshman sequence. Using a holistic rating scale would have made our task somewhat less complex since placement decisions could be made by sorting the papers into three categories, but it would not have met our broader goals.

The difficulties in using an analytic scoring method were addressed directly in the summer of 1982 with the use of computer-based processing system. The computer-based program attempted to resolve some of the problems inherent in writing sample administration as revealed by our research and by our pilot program. The processing system functioned in four major ways. The actual placement decision

was made by the computer program as it merged writing sample scores with ACT English scores. Unlike the pilot run the previous summer, the decision for course placement was given to the computer program rather than the raters or the coordinator of the training session. By freeing the raters of the responsibility for course placement, they might concentrate more directly on the analytic criteria. As designed, the computer program also allowed us the opportunity to monitor both rater and team performance in relation to ACT norms. Since our previous study of the writing sample process and rater behavior led to the conclusion that a raw score system would be inadequate, with some raters consistently scoring the same papers lower or higher than others, the decision was made to compute a T-score for each rater's work. The idea had further merit because the ACT English score could be equated on the same basis. T-scores with a mean of 50 and standard deviation of 10 became the basis of the score analysis system. Finally, the computer program allowed us the opportunity to prepare test score profiles for each student in the program. Such a profile could become a part of a student's permanent file and could be made available to the English instructor at the beginning of each term.

English Ability Test Score Profile
Date: 7-15-82

School: 1

Writing Sample

Scale	Rater A	Rater B
Mechanics	3.	4.
Sentence Structure	3.	3.
Structure of Paper	3.	3.
Development of Paper	2.	2.
Style, Clarity, Diction	2.	2.
Content and Ideas	3.	2.
Total	17.	18.
T Score 1-3	46.	42.
T Score 4-6	45.	37.
Total T Score 1-6	46.	39.

ACT SCORES

Test	Score	T Score
English	8.	33.
Mathematics	9.	42.
Social Studies	4.	33.
Natural Science	12.	36.
Composite	8.	32.

PLACEMENT RECOMMENDATION

Lab Course	***
Regular Eng. 111	___
English 112	___

Even with the program in place at the beginning of the summer, we noted how difficult a task it is for raters to consistently adhere to the analytic criteria. The tendency to read the paper for an initial impression and then score it accordingly worked against the validity of the analytical scale. Since the actual placement decision was made by numerical values given to the computer program, we were aware that some distortion in scores would occur if the analytic scale was not rigorously adhered to. This is a difficulty in all attempts at analytic scoring. Is it possible to proceed inductively from a criterion-referenced system to make placement decisions or does the process of writing evaluation negate analytic methods by creating a "halo effect" where a single trait comes to control the final evaluation? An exploratory factor analysis on over 300 themes evaluated the previous summer showed that even though six factors were listed in the analytic scale, only one factor emerged from the ratings. We reasoned that raters formed an immediate impression, decided where the student should be placed, and then scored the paper accordingly. Even though the readers made a conscious attempt to concentrate on the analytic criteria, the factor analysis showed that a single trait tended to dominate and control the placement. Since so many variables offset the writing sample evaluation, we arbitrarily weighted the writing sample 50% and ACT English 50% in order to minimize the possibility for local error in determining course placement.

In addressing the need to maintain measurement stability from session to session and from rater-group to rater-group, the computer program allowed us to make important administrative decisions as the assessment program developed. We knew from our previous investigation that standards for writing evaluation and inter-rater agreement can shift depending on the group doing the reading and the day of the evaluation. An exhausted team is likely to do a less reliable job at 4:00 p.m. on the final day of a week-long series of rating sessions than a team just beginning on a Monday afternoon. A faculty member who perceives that he or she is

apt to be "too soft" or "too tough" in applying the analytic criteria may distort the scoring and the subsequent placement decision in an effort to compensate for a perceived inaccuracy. Raters may dawdle one day when the number of examinees is light and rush another when papers arrive late and a dinner or cocktail arrangement may be pending. These and an infinite number of other distractions could cause teams to score differently from day to day and from team to team. We asked ourselves how could a monitoring system be established that not only could track rater performance but also could adjust scores when such discrepancies begin to affect the reliability of placement decisions.

One means to monitor rater performance is illustrated in the following table where correlation coefficients are given for each rater, all rater pairs, and each rater pair with the ACT.

		Rater Team Two				
Evaluator	ACT Eng.	5	6	7	8	Average
ACT English*	--	.48	.56	.49	.57	.53
Rater 5		--	.75	.47	.74	
Rater 6			--	.76	.63	
Rater 7				--	.78	
Rater 8					--	
Raters Combined						.69

*Rater correlations with ACT are the average correlations for the rater pairs. The number of writing samples for the correlations range from 97 to 117.

As may be observed on this table, the correlation with ACT range from 0.48 to 0.57, an acceptable level since the ACT English score has a proven and established success in predicting performance in writing as measured by grades. The rater intercorrelations across the six sessions ranged from 0.47 to 0.78 with the average interrater correlation being 0.69. The average correlation for the 1981 summer groups was 0.62 and for Team one, 0.54. From this we reasoned that an acceptable correlation would be around .60 if we wished to insure high reliability. If at any point team performance moves below 0.60, additional training sessions could be called to encourage

greater Consensus. We observed by such data that as raters improve their agreement with each other on the writing samples, they also improve their agreement with the ACT English measure.

We were aware that even with this monitoring system in place, undesirable fluctuations in scores may occur that would be left undetected by this procedure. Mean scores for the entire team could drift upwards or downwards without a corresponding change in the average abilities of the students tested. By carefully monitoring mean raw scores as correlated with ACT English scores, our program allowed us to determine whether day to day changes in scores were changes in the quality of the writing samples or if they were merely random fluctuations. The following table shows the mean scores for the ACT English test and the raters mean raw scores for six testing periods.

Description of ACT and Rater Means by Orientation Session Number.

Rater Team One

Orientation Session Number	Mean ACT English	Mean Raw Scores and Sample Size — Rater Number 1	2	3	4
Typical 1	15.4(115)*	17.4(54)	20.3(65)	19.8(63)	19.5(52)
Typical 2	15.8(131)	17.1(62)	21.3(65)	20.3(71)	19.9(74)
Typical 3	15.7(159)	16.3(80)	19.7(75)	18.9(74)	18.9(69)
Typical 4	15.3(120)	14.5(69)	18.6(67)	17.8(60)	18.9(62)
Average	15.6(525)	16.2(265)	20.0(272)	19.2(268)	19.3(257)
CLEP 6	21.3(82)	19.9(43)	21.8(44)	21.4(42)	20.1(45)
CLEP 7	21.7(114)	18.9(56)	18.8(60)	23.3(56)	19.8(58)
Average	21.5(196)	19.3(99)	20.1(104)	22.5(98)	19.9(103)
Maximum Variation of Means					
Sessions 1-4	0.5	2.9	2.7	2.5	1.0
Sessions 6-7	0.4	1.0	3.0	1.9	0.3

*Frequencies in Parentheses.

Two distinctly different groups were tested during these periods, one the advanced placement students as determined by ACT test scores, and the other "typical" Ferris freshmen. Both groups reported fairly homogenous test scores with the "typical group" averaging 15.6 and the advanced or CLEP group averaging 21.5. Little variation existed day by day within each group. Two problems immediately became evident as we examined this table. The first problem was correctable given the processing system. All four raters gave lower average scores to the students in the typical 4th session than they did to those in the first session even though the ACT English scores are virtually identical. Observing this drift downwards, the decision was made to adjust the cut-off scores for exemption and 111 placement somewhat higher to compensate for the vacillation. While it would be desirable to have stability in rating patterns from day to day, when the mean ACT score indicates a major change in expected writing performance, then the raters' average scores would change also. As the table illustrates, two raters provided advanced placement group averages that were 3 or more points higher than the "typical" group average, while the other raters provided advanced placement group averages that were only 0.6 and 0.1 points higher. This data can alert the test administrators to the possibility that the two raters who were not responsive to changes in group ability may be evaluating the papers according to the day's norm rather than to the agreed-upon analytical criteria. If these two raters scored the same student papers, it is probable that they would receive a lower course placement than would be given by the other raters. Such discrepancies could best be resolved by bringing the team together for further training sessions. Not to do so would create misplacement that could affect many students. Since, at Ferris, student ability levels are not uniform across the summer orientation sessions, raters should avoid the adoption of holistic patterns of scoring where standards are relative to the group being tested. An excellent understanding of the criterion-referenced scale is essential in this entire process.

After administering the writing sample to over 1,300 students during the summer, 1982 orientation program, we were determined to follow-up student performance in the various English courses during the 1982-83 academic year. As illustrated in the following table, we were aware that based on the testing procedure that was followed, approximately a quarter of the students would have been placed differently using either the writing sample or the ACT score used alone. During a summer orientation period where approximately 3,000 students would be advised and placed in various courses, over 750 students would have been placed differently. Whether the different placement would have improved placement accuracy is another question

English Course Placement ACT-WS (50-50)

ACT SCORES	ENG LAB	ENG 111	ENG 112	PROBABILITY FOR 111 EXEMPTION GIVEN ACT SCORE C-B-E	
32			2	100	
:					
:			31	100	
25			9	100	
24		4	37	90	
23		17	40	70	
22		37	70	65 ↑	(possible ACT
21	1	47	49	51	score for 111 placement)
20		52	26	33	
19	1	72	21	22	
18	2	69	8	10	
17	1	76	11	13	
16	5	76	3	04	
15	9	108	8	06	
14	14	83	5	05	
13	16	64	3	04	(former score for
12	44	33	1	01 ↓	lab placement)
11	36	25			
10	33	17			
9	23	1			
8	31				
:					
:	17				
4	3				
TOTAL	236	781	324		

The follow-up study asked certain questions crucial to the implementation of the testing project. Would students be misplaced in more advanced levels of a course, thereby increasing the risk of failure in the freshman sequence? Does the writing sample-ACT score procedure improve on the ACT procedure used alone as a way of placing students in appropriate courses in the curriculum? How closely do ACT English scores relate to performance as measured by grades in a freshman writing sequence? While these questions may seem minor in comparison to the broader goals which involve curriculum, teaching attitude and performance, and political effects across the campus, they are essential in insuring that the testing procedure yields valid and reliable data that can be depended on in the placement of students.

As shown in the following table, for those students exempted from the 111 course and placed directly into 112, only five or 3.7 received D and F grades, as compared with 20% for all students registered in English 112 in the fall, 1982 term. Given the range of student motivation and other variables that determine success in courses, it seems unlikely that all five failed to receive a C or better due to lack of ability.

FREQUENCY DISTRIBUTIONS FOR ENGLISH 112 FINAL GRADES
CATEGORIZED ON ACT TEST SCORES

	ACT English Score Intervals			
Final Grades	25-Plus	24-22	21-15	14-13
A	4	40	20	
B	1	17	26	
C	1	6	13	2
D		1	2	
F		1	1	
N=	6	65	62	2
GPA=	3.35	3.35	2.96	2.00
D-F Rate=	0%	3.1%	4.5%	0%
Overall			4.5%	0%
N=	135			
GPA=	3.15			

Of the 565 students placed in the regular 111, 42 students would have been placed in the writing lab if an ACT cut-off of 12 were to be used for lab placement. Of those only 3 or 7.1% were not successful in doing C work, which would tend to support the continued use of the sample for all students. However, 523 students who scored 13 and above would have been placed in the 111 without the use of the sample; thus, the sample could be viewed as a redundant measure if course placement remains the main objective for a writing sample program.

In summary, by monitoring inter-group and individual performance, it is possible to accomplish the following goals by means of a computer-based processing system:

1. Cut-off scores for exemption and placement can be adjusted if teams begin to show vacillation that might seriously affect placement decisions.

2. By continual monitoring of interrater agreement, training sessions can be called when raters begin to show signs of falling away from acceptable standards of performance.

3. Individual rater scores can be adjusted if marked changes occur in rating patterns. Once a rater is "fixed" in relation to a group norm, marked divergence from his rating pattern may indicate that problems or distractions have prevented the rater from performing at his optimal level. Since the reason for vacillating scores are infinite, computer monitoring can assist the program administrator in detecting where distractions occur that ultimately affect performance. Some of these problems can be corrected, such as not interrupting a faculty member to answer a phone, to give a guest lecture in someone's class, or to ask for a favor the upcoming fall term. Others, more major, can be dealt with given the documentation provided by the program.

In the minds of most members of the English profession, placement decisions cannot rest solely on scores gathered from national testing practices of the SAT and the ACT. The credibility of their assessment tools, since they rely almost exclusively on indirect measurements of ability, is not likely to improve now or in the near future. On the other hand, for those specialists in educational measurement and for those in the highest levels of higher educational administration writing samples will not gain easy acceptance since for all the effort and time expended, no studies have as yet appeared revealing any significant increase in placement accuracy.

The value of such an assessment procedure rests in its placement within a total package of activities designed to provide program and faculty development. If it is true that testing "drives the curriculum" then a program in writing sample measurement can exact a powerful and positive influence on curricular development. However, to gloss over the difficulties of administering the sample can not only create problems in proper course placement and fairness to students, but it also might negate the desired side effects which in the long run can yield the most rewards. It is our belief that a computer-assisted monitoring system for writing sample testing can offer solutions to problems that have long eluded the profession in its attempt to match and better the success of standardized objective tests.

FOOTNOTES

[1]Our ongoing study has not contradicted what Robert Ebel has asserted in Practical Problems in Educational Measurement (1980), that "there is no credible evidence that essay tests provide reliable measures of complex skills, abstract abilities, or higher mental processes that are beyond the scope of objective tests." (125)

[2]H.M. Breland and J. L. Gaynor, "A Comparison of Direct and Indirect Assessments of Writing Skill," Journal of Educational Measurement, 1979, 16, 119-128.

[3]David C. Hughes, Brian Keeling, and Bryan T. Tuck, "The Influence of Context Position and Scoring Method on Essay Scoring," Journal of Educational Measurement, 1980, 17:2, p. 131.

[4]Steele, Joe M. "The Assessment of Writing Via Qualitative Ratings of Writing Samples." Paper presented at the Annual Meeting of the National Council on Measurement in Education in San Francisco, April 9, 1979.

[5]Clinton I. Chase, "The Impact of Achievement Expectations and Handwriting Quality on Scoring Essay Tests," Journal of Educational Measurement, 1979, 16:1, 34-42. J.C. Marshall, "Writing Neatness, Composition Errors, and Essay Grades Reexamined," Journal of Educational Research, 1979, 65, 213-215.

[6]A similar conclusion was drawn by Sarah Warshaver Freeman by carrying out a factor analysis on a study of an analytic scoring procedure used with college students in California. Her study suggests that "holistic and analytic scales measured one trait of compositions" (252) with usage tending to dominate over other categories such as voice, development, organization, word choice and sentence structure. "Influence on Evaluators of Expository Essays: Beyond the Text," Research in the Teaching of English, 1981, 15:3, 245-255.